EOSDIS
NASA'S EARTH OBSERVING SYSTEM
DATA AND INFORMATION SYSTEM

International Metadata Standards and Enterprise Data Quality Metadata Systems

IN027 - December 15, 2016

Ted Habermann
Director of Earth Science
The HDF Group
thabermann@hdfgroup.org

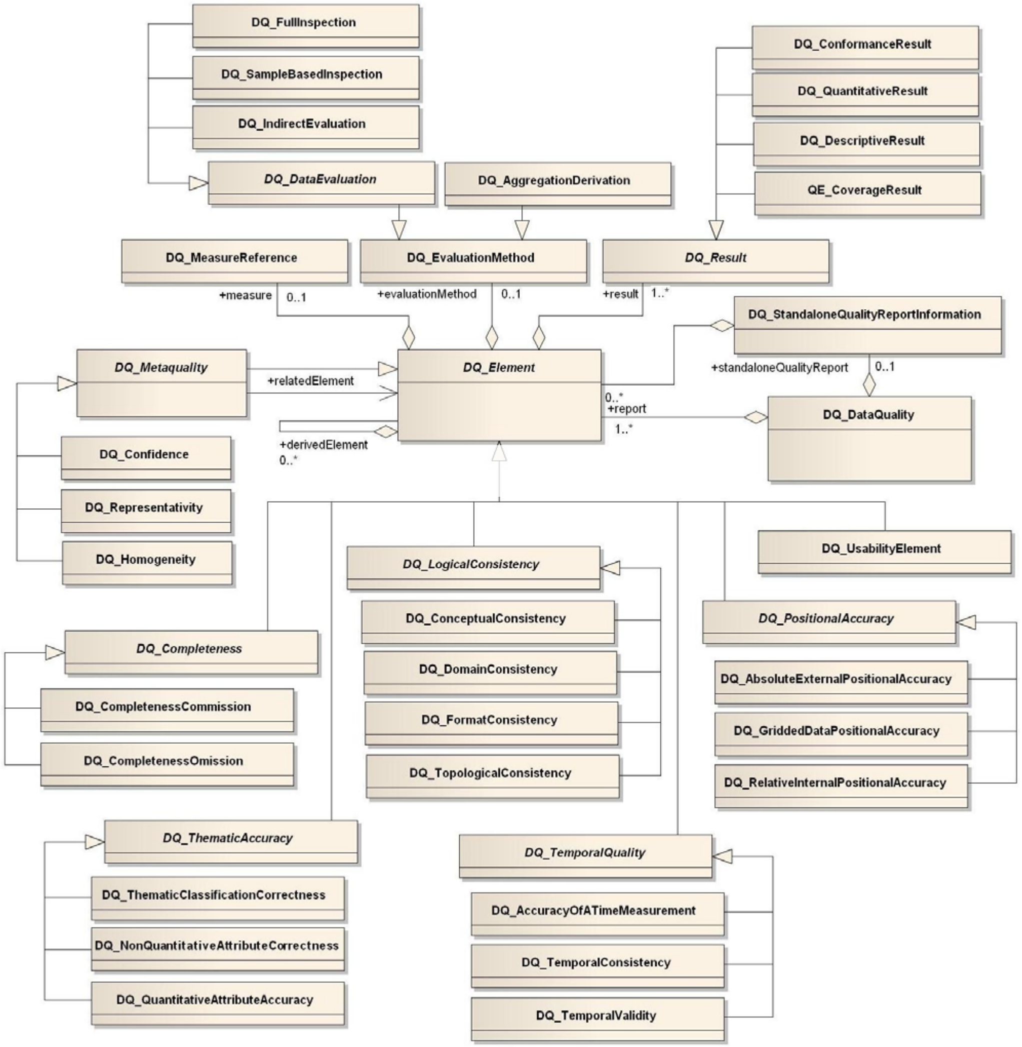
The HDF Group

This work was supported by NASA/GSFC under
Raytheon Co. contract number NNG15HZ39C

The Big Picture

ISO 19157 is a conceptual model of data quality metadata that was recently approved as an international standard. It combines three older standards into a unified model for describing data quality.

Many of the principle elements of this conceptual model are abstract, and can be implemented in several ways.



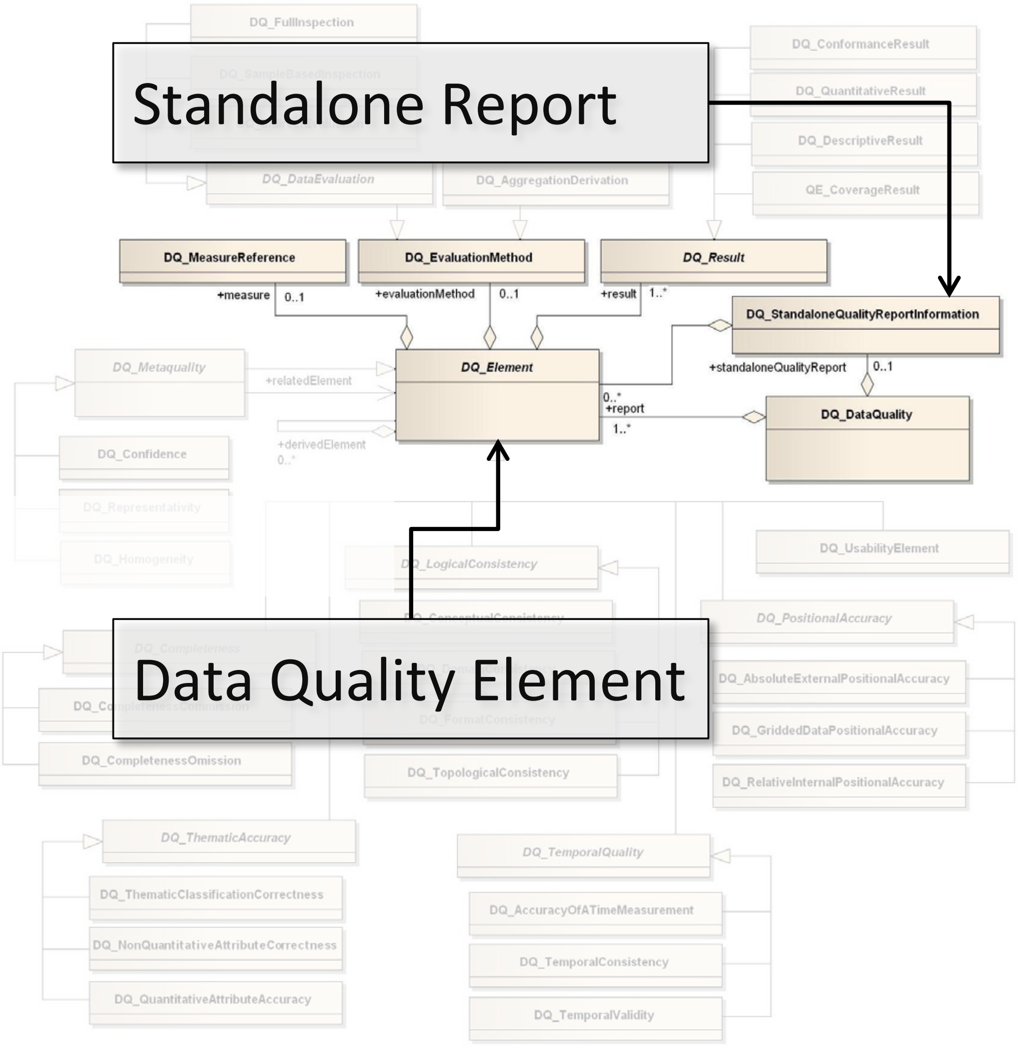
DQ_FullInspection
DQ_SampleBasedInspection
DQ_IndirectEvaluation
DQ_DataEvaluation
DQ_AggregationDerivation
DQ_ConformanceResult
DQ_QuantitativeResult
DQ_DescriptiveResult
QE_CoverageResult
DQ_MeasureReference
DQ_EvaluationMethod
DQ_Result
+measure 0..1
+evaluationMethod 0..1
+result 1..*
DQ_StandaloneQualityReportInformation
DQ_Metaquality
+relatedElement
DQ_Element
+standaloneQualityReport 0..1
0..* +report 1..*
DQ_DataQuality
+derivedElement 0..*
DQ_Confidence
DQ_Representativity
DQ_Homogeneity
DQ_UsabilityElement
DQ_LogicalConsistency
DQ_ConceptualConsistency
DQ_DomainConsistency
DQ_FormatConsistency
DQ_TopologicalConsistency
DQ_PositionalAccuracy
DQ_AbsoluteExternalPositionalAccuracy
DQ_GriddedDataPositionalAccuracy
DQ_RelativeInternalPositionalAccuracy
DQ_Completeness
DQ_CompletenessCommission
DQ_CompletenessOmission
DQ_ThematicAccuracy
DQ_ThematicClassificationCorrectness
DQ_NonQuantitativeAttributeCorrectness
DQ_QuantitativeAttributeAccuracy
DQ_TemporalQuality
DQ_AccuracyOfATimeMeasurement
DQ_TemporalConsistency
DQ_TemporalValidity

The Big Picture

ISO 19157 is a conceptual model of data quality metadata that was recently approved as an international standard. It combines three older standards into a unified model for describing data quality.

Many of the principle elements of this conceptual model are abstract, they can be implemented in several ways.

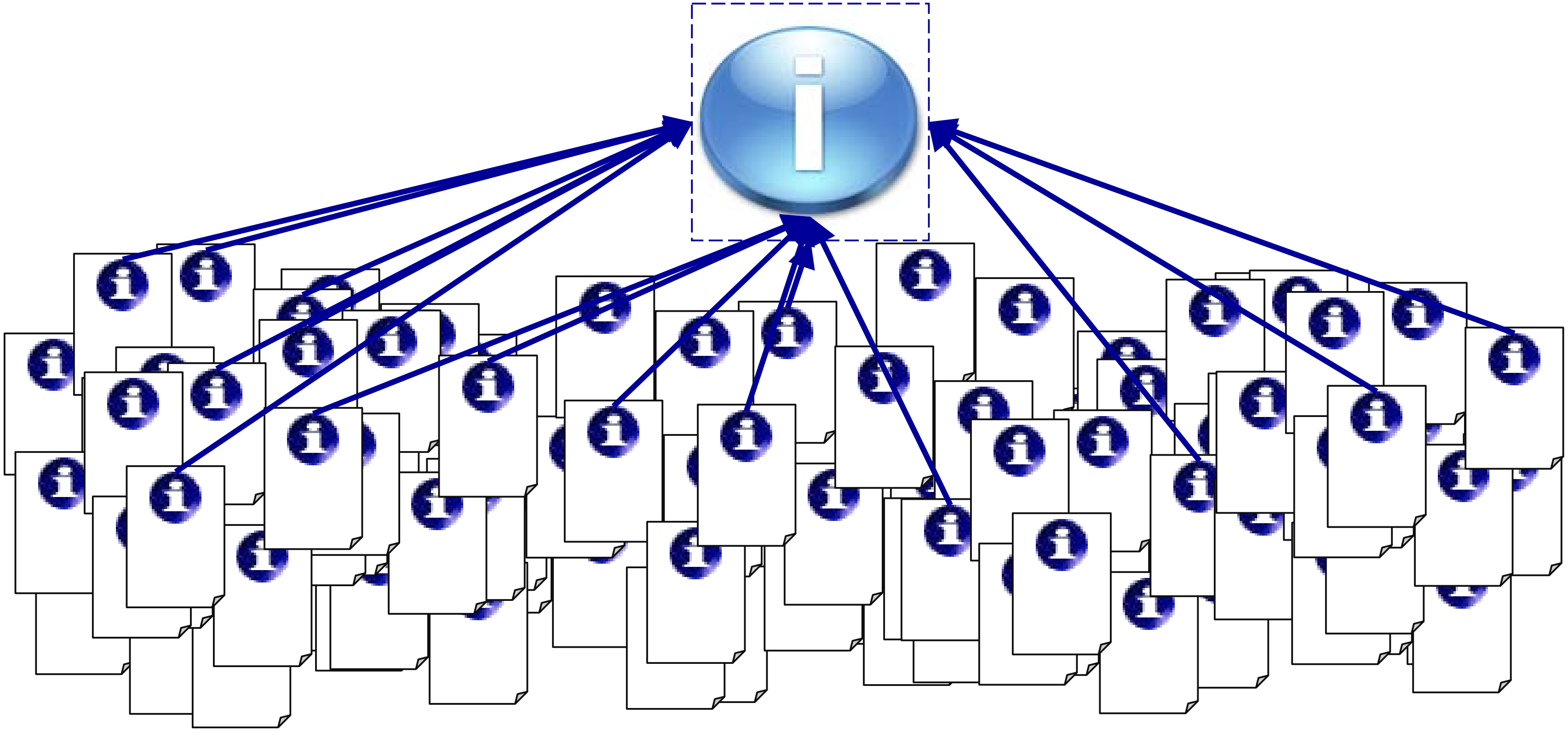
When only the abstract concepts are considered, the model is very simple.



Standalone Report

Data Quality Element

Enterprise Systems?



Stand Alone Quality Reports

“There are papers and web pages that describe the quality of my data.”

Papers and reports that describe data quality are StandAloneReports. Metadata can include brief descriptions of the results (abstracts) and references to any number of these (citations).

DQ_StandaloneQualityReportInformation

+ abstract : CharacterString
+ reportReference: CI_Citation

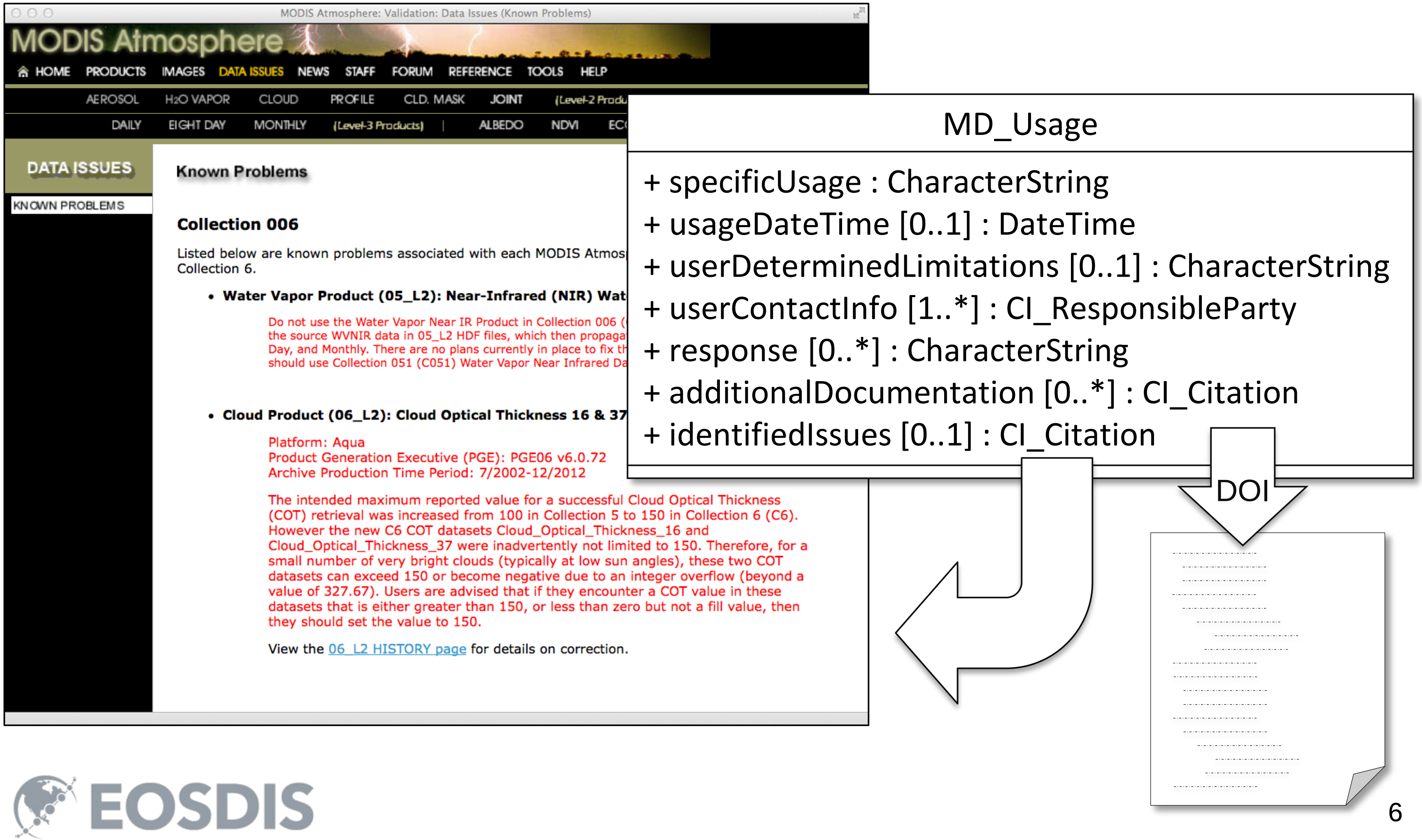
Abstract: The fire training-set may also have been biased against savanna and savanna woodland fires since their detection is more difficult than in humid, forest environments with cool background temperatures [Malingreau, 1990]. There may, therefore, be an under-sampling of warmer background environments.

DOI

Citation: Malingreau J.P, 1990, The contribution of remote sensing to the global monitoring of fires in tropical and subtropical ecosystems. In: *Fire in Tropical Biota*, (J.G. Goldammer , editor), Springer Verlag , Berlin: 337-370.

Data Usage (19115-1)

“Users increase our understanding of data quality. We need to keep them in the loop.”


MODIS Atmosphere: Validation: Data Issues (Known Problems)
MODIS Atmosphere
HOME PRODUCTS IMAGES DATA ISSUES NEWS STAFF FORUM REFERENCE TOOLS HELP
AEROSOL H2O VAPOR CLOUD PROFILE CLD. MASK JOINT (Level-2 Produ
DAILY EIGHT DAY MONTHLY (Level-3 Products) | ALBEDO NDVI EC
DATA ISSUES
KNOWN PROBLEMS
DOI

Known Problems

Collection 006

Listed below are known problems associated with each MODIS Atmos Collection 6.

- **Water Vapor Product (05_L2): Near-Infrared (NIR) Wat**

  Do not use the Water Vapor Near IR Product in Collection 006 (the source WVNIR data in 05_L2 HDF files, which then propaga Day, and Monthly. There are no plans currently in place to fix th should use Collection 051 (C051) Water Vapor Near Infrared Da

- **Cloud Product (06_L2): Cloud Optical Thickness 16 & 37**

  Platform: Aqua
  Product Generation Executive (PGE): PGE06 v6.0.72
  Archive Production Time Period: 7/2002-12/2012

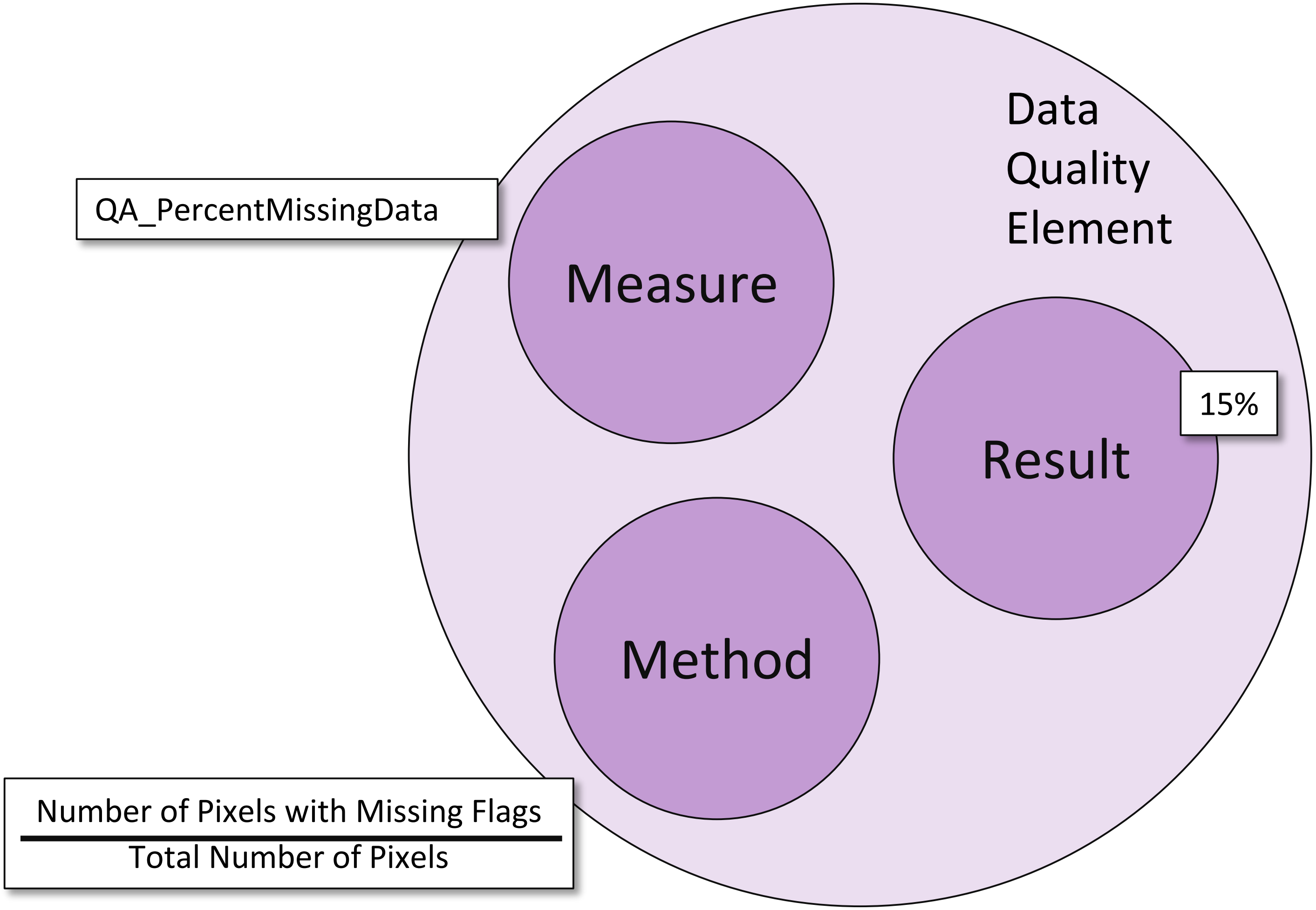
  The intended maximum reported value for a successful Cloud Optical Thickness (COT) retrieval was increased from 100 in Collection 5 to 150 in Collection 6 (C6). However the new C6 COT datasets Cloud_Optical_Thickness_16 and Cloud_Optical_Thickness_37 were inadvertently not limited to 150. Therefore, for a small number of very bright clouds (typically at low sun angles), these two COT datasets can exceed 150 or become negative due to an integer overflow (beyond a value of 327.67). Users are advised that if they encounter a COT value in these datasets that is either greater than 150, or less than zero but not a fill value, then they should set the value to 150.

  View the 06_L2 HISTORY page for details on correction.

MD_Usage

+ specificUsage : CharacterString
+ usageDateTime [0..1] : DateTime
+ userDeterminedLimitations [0..1] : CharacterString
+ userContactInfo [1..*] : CI_ResponsibleParty
+ response [0..*] : CharacterString
+ additionalDocumentation [0..*] : CI_Citation
+ identifiedIssues [0..1] : CI_Citation

What is a Data Quality Element?



Data Quality Element

Measure

Result

Method

QA_PercentMissingData

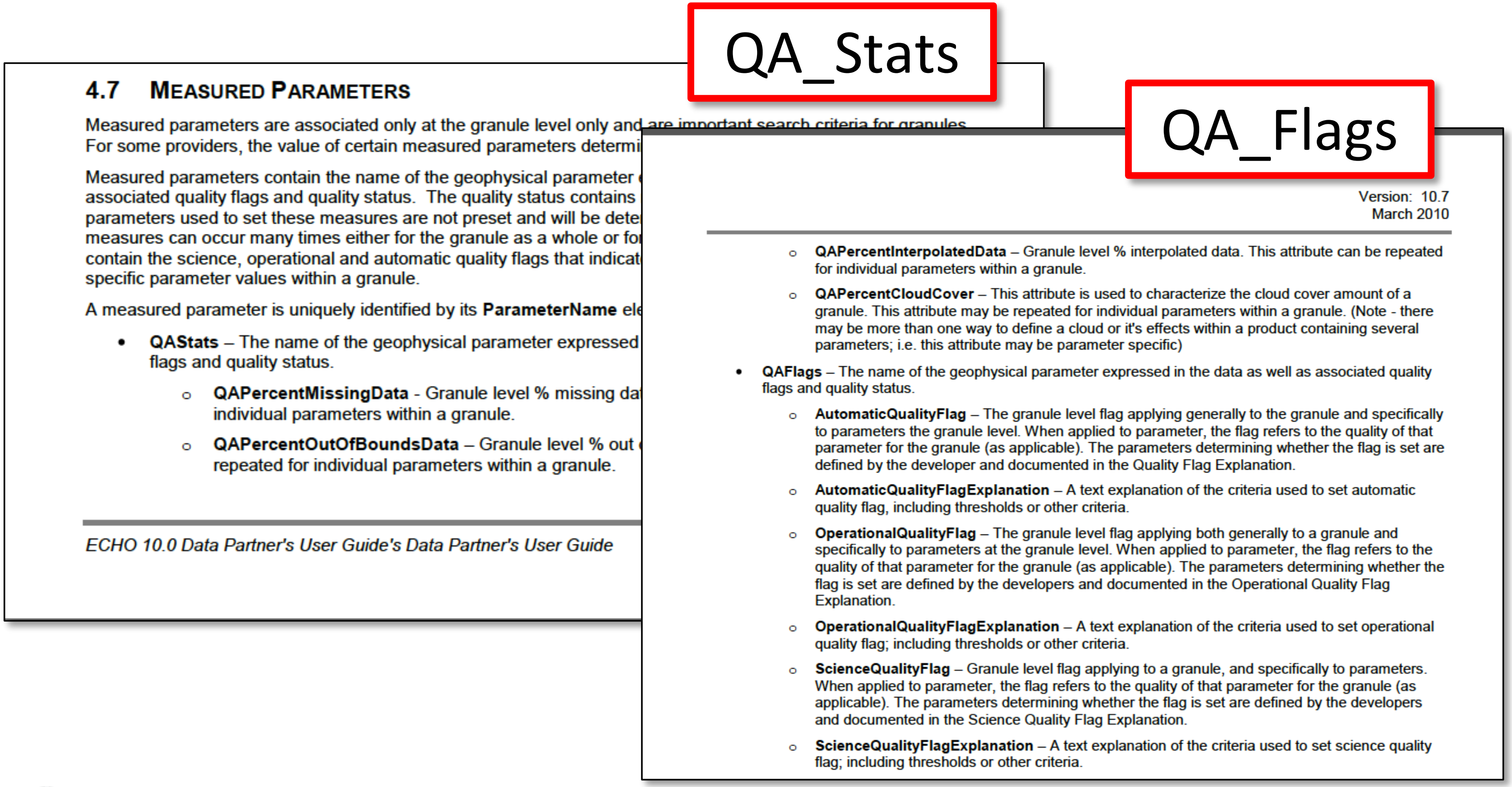
15%

$$\frac{\text{Number of Pixels with Missing Flags}}{\text{Total Number of Pixels}}$$

What Are Quality Measures?

"My metadata already include data quality measures ."

NASA EOSDIS metadata includes two types of quality measures.



QA_Stats

QA_Flags

4.7 MEASURED PARAMETERS

Measured parameters are associated only at the granule level only and are important search criteria for granules.
For some providers, the value of certain measured parameters determi

Measured parameters contain the name of the geophysical parameter
associated quality flags and quality status. The quality status contains
parameters used to set these measures are not preset and will be dete
measures can occur many times either for the granule as a whole or fo
contain the science, operational and automatic quality flags that indicat
specific parameter values within a granule.

A measured parameter is uniquely identified by its **ParameterName** el

- **QAStats** – The name of the geophysical parameter expressed
  flags and quality status.
  - **QAPercentMissingData** - Granule level % missing da
    individual parameters within a granule.
  - **QAPercentOutOfBoundsData** – Granule level % out
    repeated for individual parameters within a granule.

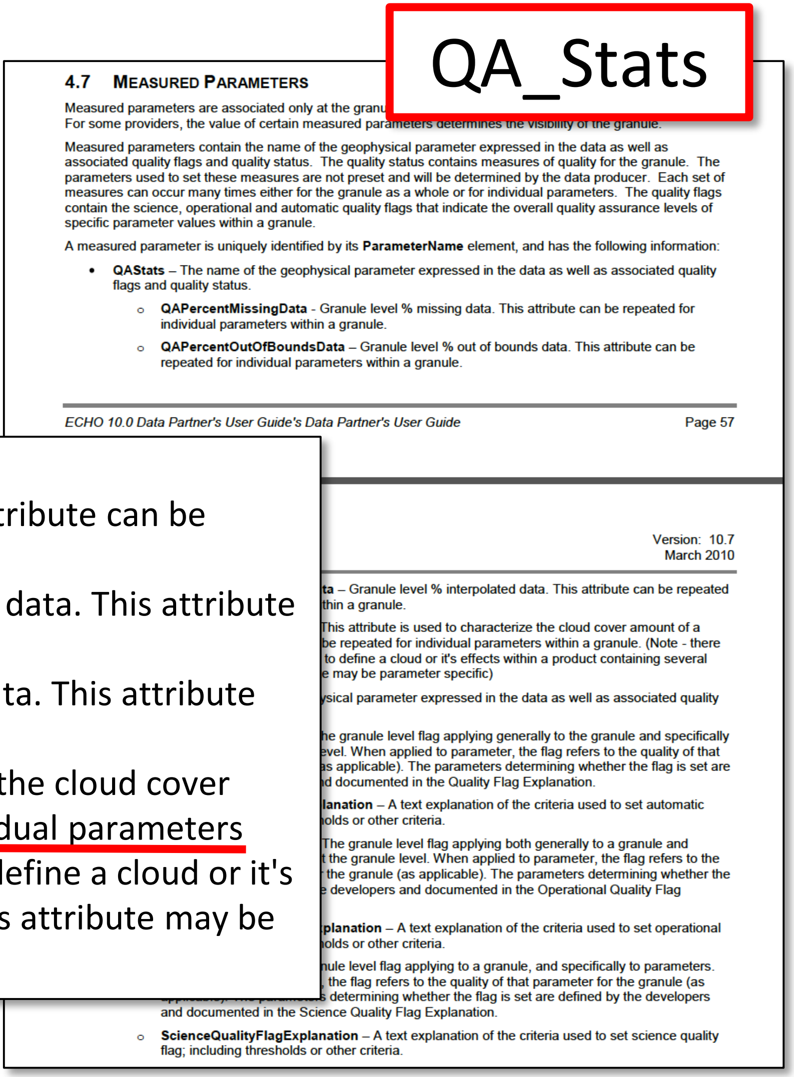
ECHO 10.0 Data Partner's User Guide's Data Partner's User Guide

Version: 10.7
March 2010

  - **QAPercentInterpolatedData** – Granule level % interpolated data. This attribute can be repeated for individual parameters within a granule.
  - **QAPercentCloudCover** – This attribute is used to characterize the cloud cover amount of a granule. This attribute may be repeated for individual parameters within a granule. (Note - there may be more than one way to define a cloud or it's effects within a product containing several parameters; i.e. this attribute may be parameter specific)
- **QAFlags** – The name of the geophysical parameter expressed in the data as well as associated quality flags and quality status.
  - **AutomaticQualityFlag** – The granule level flag applying generally to the granule and specifically to parameters the granule level. When applied to parameter, the flag refers to the quality of that parameter for the granule (as applicable). The parameters determining whether the flag is set are defined by the developer and documented in the Quality Flag Explanation.
  - **AutomaticQualityFlagExplanation** – A text explanation of the criteria used to set automatic quality flag, including thresholds or other criteria.
  - **OperationalQualityFlag** – The granule level flag applying both generally to a granule and specifically to parameters at the granule level. When applied to parameter, the flag refers to the quality of that parameter for the granule (as applicable). The parameters determining whether the flag is set are defined by the developers and documented in the Operational Quality Flag Explanation.
  - **OperationalQualityFlagExplanation** – A text explanation of the criteria used to set operational quality flag; including thresholds or other criteria.
  - **ScienceQualityFlag** – Granule level flag applying to a granule, and specifically to parameters. When applied to parameter, the flag refers to the quality of that parameter for the granule (as applicable). The parameters determining whether the flag is set are defined by the developers and documented in the Science Quality Flag Explanation.
  - **ScienceQualityFlagExplanation** – A text explanation of the criteria used to set science quality flag; including thresholds or other criteria.

What Are Quality Measures?

“I use consistent Quality Measures across many products.”



QA_Stats

4.7 Measured Parameters

Measured parameters are associated only at the granu
For some providers, the value of certain measured parameters determines the visibility of the granule.

Measured parameters contain the name of the geophysical parameter expressed in the data as well as associated quality flags and quality status. The quality status contains measures of quality for the granule. The parameters used to set these measures are not preset and will be determined by the data producer. Each set of measures can occur many times either for the granule as a whole or for individual parameters. The quality flags contain the science, operational and automatic quality flags that indicate the overall quality assurance levels of specific parameter values within a granule.

A measured parameter is uniquely identified by its **ParameterName** element, and has the following information:

- **QAStats** – The name of the geophysical parameter expressed in the data as well as associated quality flags and quality status.
  - ○ **QAPercentMissingData** - Granule level % missing data. This attribute can be repeated for individual parameters within a granule.
  - ○ **QAPercentOutOfBoundsData** – Granule level % out of bounds data. This attribute can be repeated for individual parameters within a granule.

ECHO 10.0 Data Partner's User Guide's Data Partner's User Guide Page 57

Version: 10.7
March 2010

ta – Granule level % interpolated data. This attribute can be repeated
thin a granule.

This attribute is used to characterize the cloud cover amount of a
be repeated for individual parameters within a granule. (Note - there
to define a cloud or it's effects within a product containing several
e may be parameter specific)

ysical parameter expressed in the data as well as associated quality

he granule level flag applying generally to the granule and specifically
evel. When applied to parameter, the flag refers to the quality of that
as applicable). The parameters determining whether the flag is set are
d documented in the Quality Flag Explanation.

lanation – A text explanation of the criteria used to set automatic
olds or other criteria.

The granule level flag applying both generally to a granule and
t the granule level. When applied to parameter, the flag refers to the
the granule (as applicable). The parameters determining whether the
e developers and documented in the Operational Quality Flag

planation – A text explanation of the criteria used to set operational
olds or other criteria.

nule level flag applying to a granule, and specifically to parameters.
, the flag refers to the quality of that parameter for the granule (as
s determining whether the flag is set are defined by the developers
and documented in the Science Quality Flag Explanation.

○ **ScienceQualityFlagExplanation** – A text explanation of the criteria used to set science quality flag; including thresholds or other criteria.

QAStats – **Standard measures for all products**

QAPercentMissingData - Granule level % missing data. This attribute can be repeated for individual parameters within a granule.

QAPercentOutOfBoundsData – Granule level % out of bounds data. This attribute can be repeated for individual parameters within a granule.

QAPercentInterpolatedData – Granule level % interpolated data. This attribute can be repeated for individual parameters within a granule.

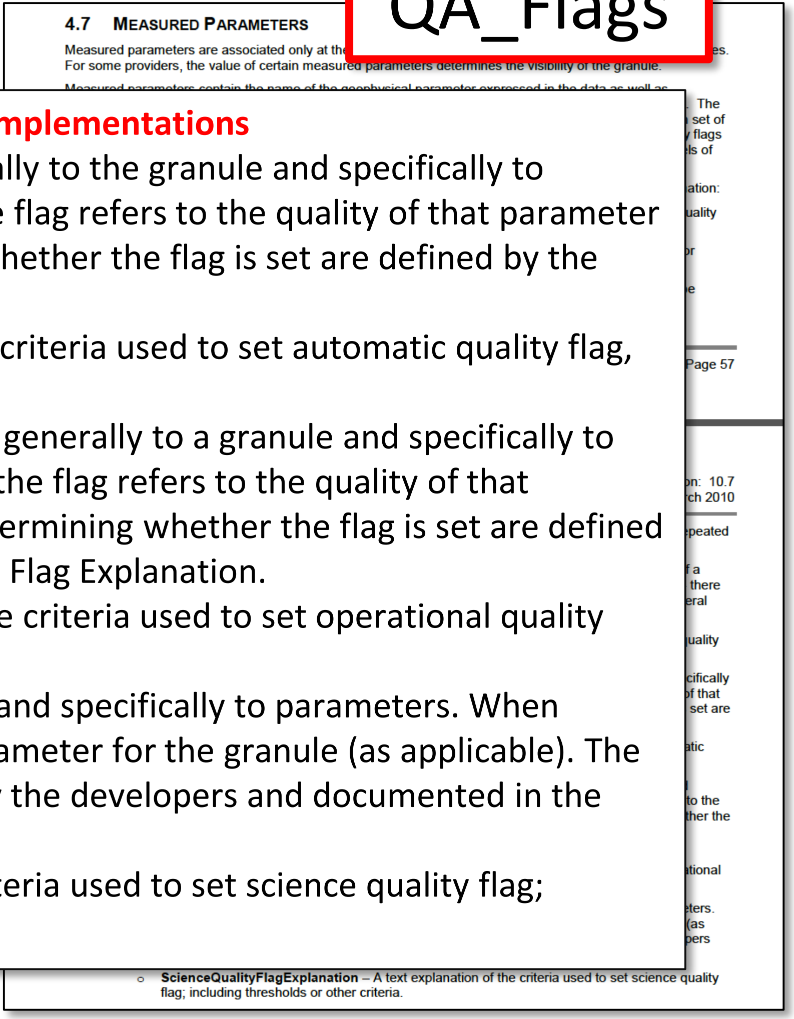
QAPercentCloudCover – This attribute is used to characterize the cloud cover amount of a granule. This attribute may be repeated for individual parameters within a granule. (Note - there may be more than one way to define a cloud or it's effects within a product containing several parameters; i.e. this attribute may be parameter specific)

EOSDIS

What Are Quality Measures?

"I use consistent types of Quality Measure across many products."

QA_Flags



4.7 MEASURED PARAMETERS

Measured parameters are associated only at th… es.
For some providers, the value of certain measured parameters determines the visibility of the granule.

Measured parameters contain the name of the geophysical parameter expressed in the data as well as

The
set of
flags
ls of
ation:
uality
r
e

Page 57

n: 10.7
ch 2010

peated
f a
there
eral
uality
cifically
f that
set are
atic
to the
her the
tional
ters.
(as
pers

○ **ScienceQualityFlagExplanation** – A text explanation of the criteria used to set science quality flag; including thresholds or other criteria.

QAFlags – **Classes of quality measures with product specific implementations**

AutomaticQualityFlag – The granule level flag applying generally to the granule and specifically to parameters the granule level. When applied to parameter, the flag refers to the quality of that parameter for the granule (as applicable). The parameters determining whether the flag is set are defined by the developer and documented in the Quality Flag Explanation.

AutomaticQualityFlagExplanation – A text explanation of the criteria used to set automatic quality flag, including thresholds or other criteria.

OperationalQualityFlag – The granule level flag applying both generally to a granule and specifically to parameters at the granule level. When applied to parameter, the flag refers to the quality of that parameter for the granule (as applicable). The parameters determining whether the flag is set are defined by the developers and documented in the Operational Quality Flag Explanation.

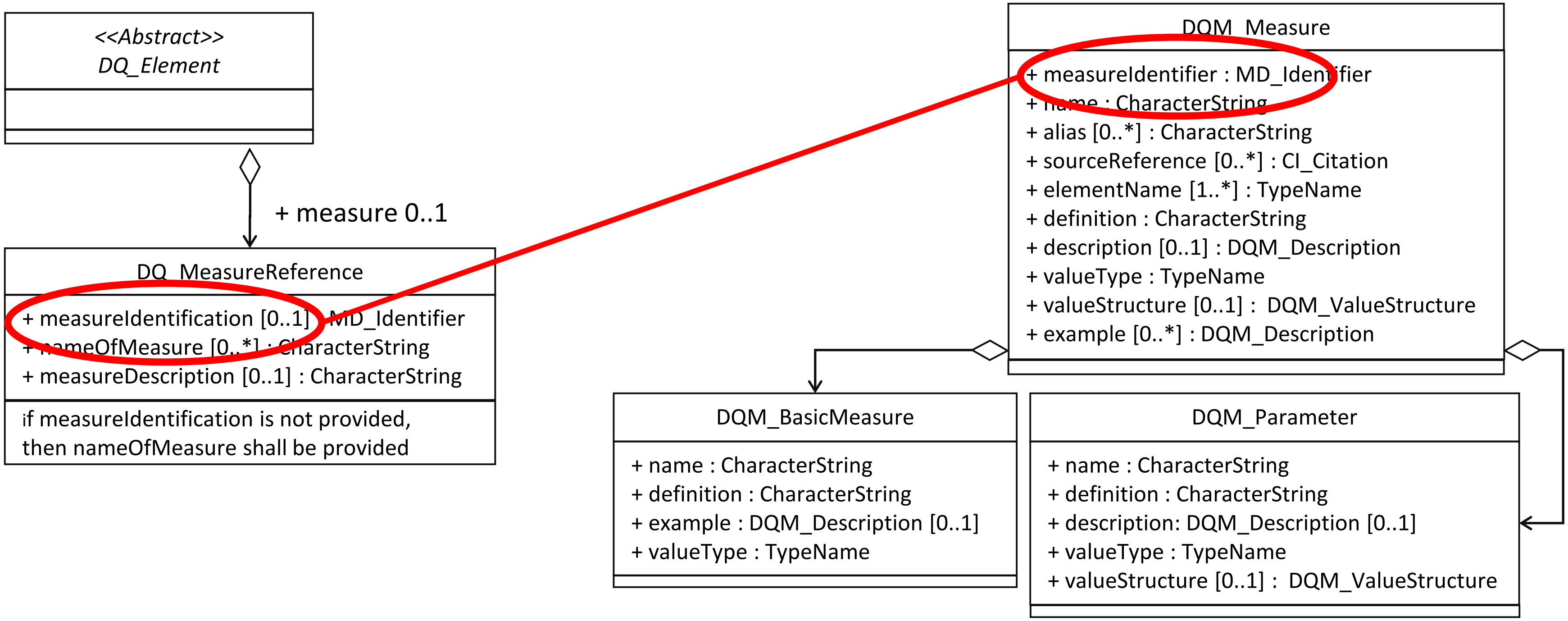
OperationalQualityFlagExplanation – A text explanation of the criteria used to set operational quality flag; including thresholds or other criteria.

ScienceQualityFlag – Granule level flag applying to a granule, and specifically to parameters. When applied to parameter, the flag refers to the quality of that parameter for the granule (as applicable). The parameters determining whether the flag is set are defined by the developers and documented in the Science Quality Flag Explanation.

ScienceQualityFlagExplanation – A text explanation of the criteria used to set science quality flag; including thresholds or other criteria.

Data Quality Measures

"My data quality measures are consistently described in a database ."



<<Abstract>>
DQ_Element

+ measure 0..1

DQ_MeasureReference

+ measureIdentification [0..1] : MD_Identifier
+ nameOfMeasure [0..*] : CharacterString
+ measureDescription [0..1] : CharacterString

if measureIdentification is not provided, then nameOfMeasure shall be provided

DQM_Measure

+ measureIdentifier : MD_Identifier
+ name : CharacterString
+ alias [0..*] : CharacterString
+ sourceReference [0..*] : CI_Citation
+ elementName [1..*] : TypeName
+ definition : CharacterString
+ description [0..1] : DQM_Description
+ valueType : TypeName
+ valueStructure [0..1] : DQM_ValueStructure
+ example [0..*] : DQM_Description

DQM_BasicMeasure

+ name : CharacterString
+ definition : CharacterString
+ example : DQM_Description [0..1]
+ valueType : TypeName

DQM_Parameter

+ name : CharacterString
+ definition : CharacterString
+ description: DQM_Description [0..1]
+ valueType : TypeName
+ valueStructure [0..1] : DQM_ValueStructure

ISO 19157 includes a DQ_MeasureReference designed to provide a connection to a detailed description of the quality measure.

DQM_Description

+ textDescription: CharacterString
+ extendedDescription [0..1] : MD_BrowseGraphic

Data Quality Measures

“I need to clearly and consistently explain how I measure quality.”

The ISO model for quality measures includes identifiers, definitions, descriptions, references and illustrations.

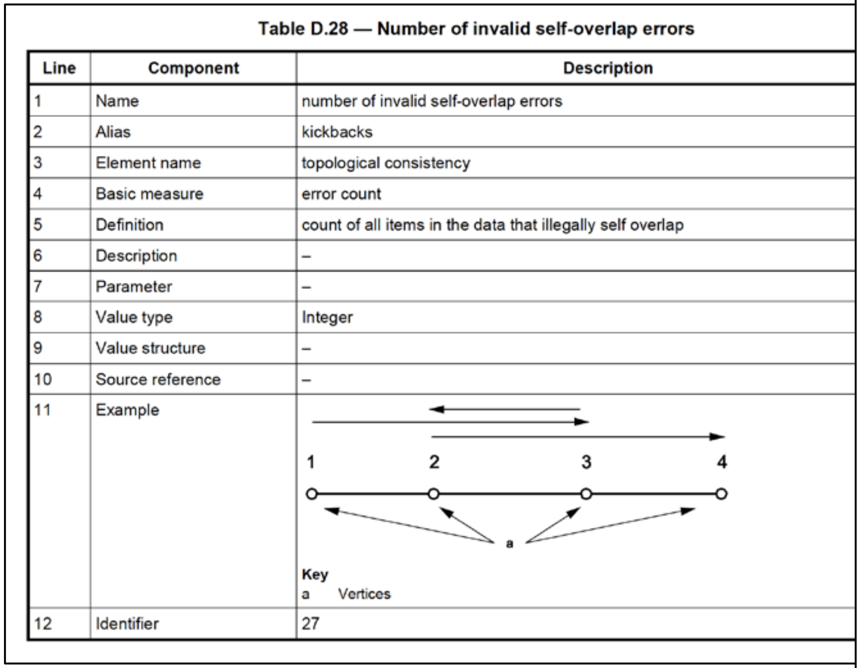


Table D.28 — Number of invalid self-overlap errors

Line	Component	Description
1	Name	number of invalid self-overlap errors
2	Alias	kickbacks
3	Element name	topological consistency
4	Basic measure	error count
5	Definition	count of all items in the data that illegally self overlap
6	Description	–
7	Parameter	–
8	Value type	Integer
9	Value structure	–
10	Source reference	–
11	Example	1 2 3 4 a **Key** a Vertices
12	Identifier	27

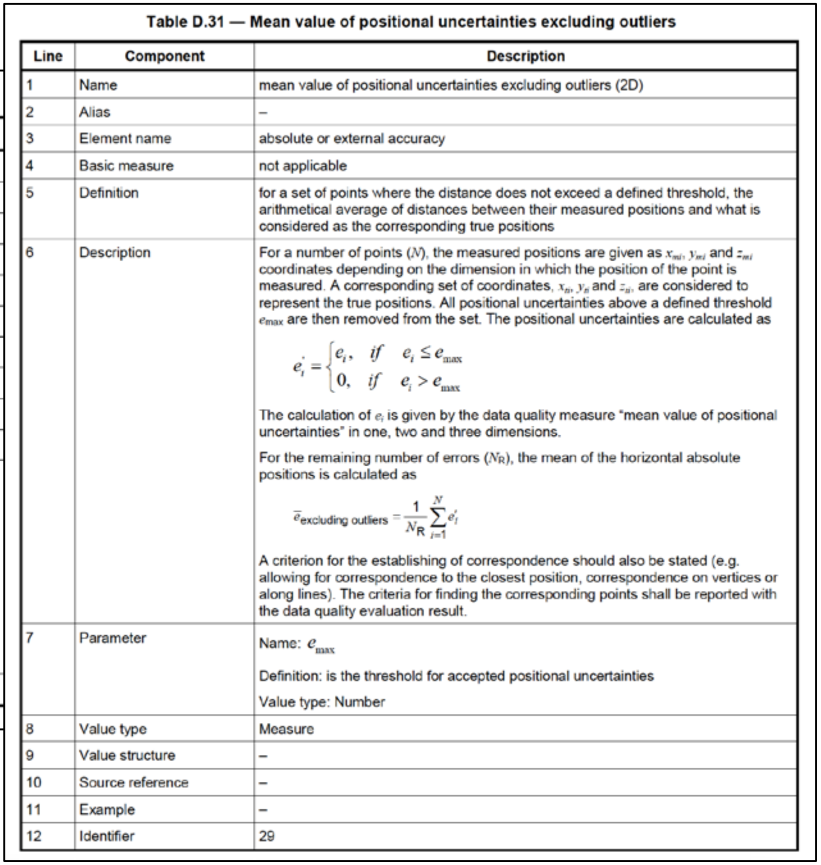
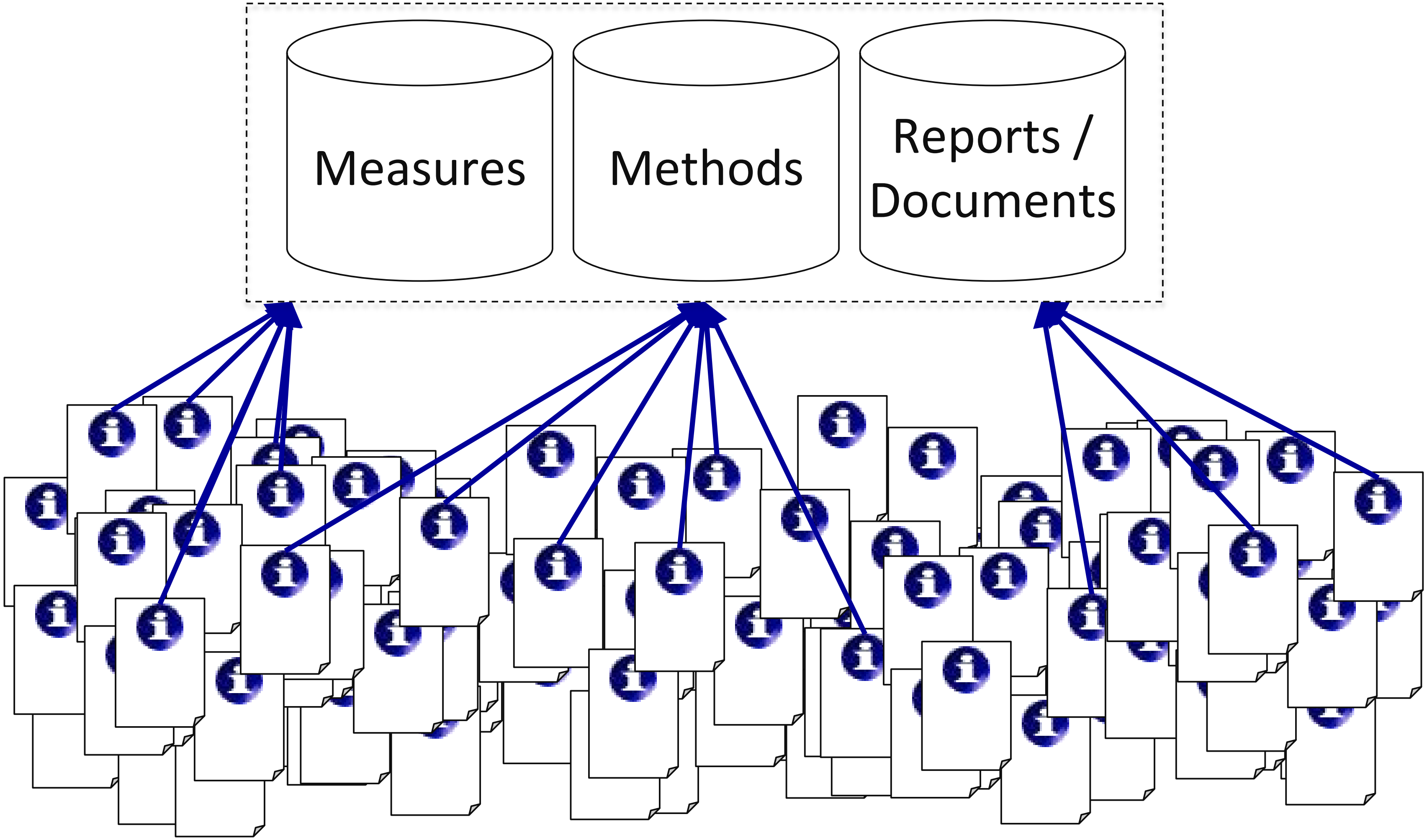


Table D.31 — Mean value of positional uncertainties excluding outliers

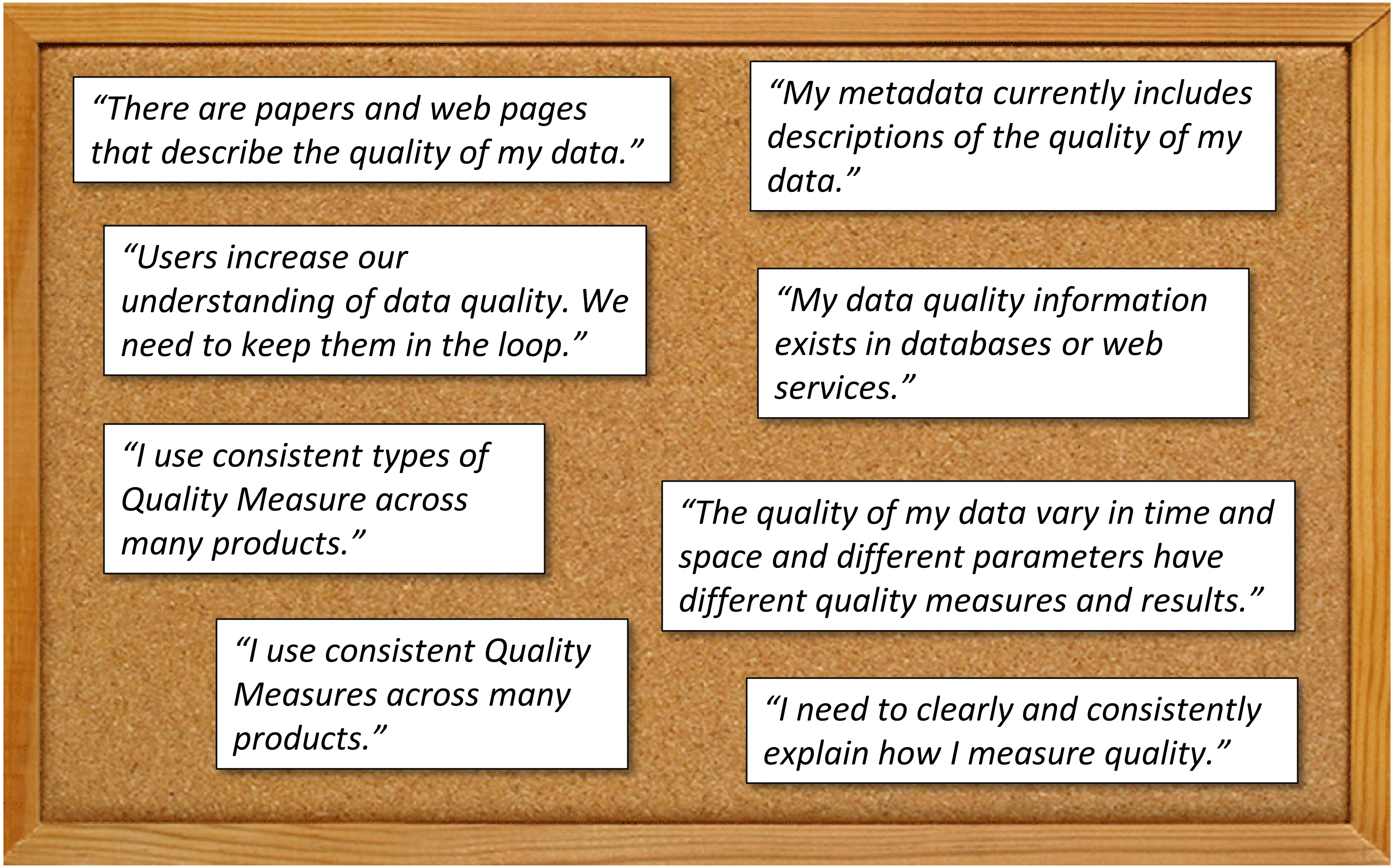
Line	Component	Description
1	Name	mean value of positional uncertainties excluding outliers (2D)
2	Alias	–
3	Element name	absolute or external accuracy
4	Basic measure	not applicable
5	Definition	for a set of points where the distance does not exceed a defined threshold, the arithmetical average of distances between their measured positions and what is considered as the corresponding true positions
6	Description	For a number of points (N), the measured positions are given as x_{mi}, y_{mi} and z_{mi} coordinates depending on the dimension in which the position of the point is measured. A corresponding set of coordinates, x_{ti}, y_{ti} and z_{ti}, are considered to represent the true positions. All positional uncertainties above a defined threshold $e_{\max}$ are then removed from the set. The positional uncertainties are calculated as $e_i' = \begin{cases} e_i, & if \quad e_i \le e_{\max} \\ 0, & if \quad e_i > e_{\max} \end{cases}$ The calculation of e_i is given by the data quality measure "mean value of positional uncertainties" in one, two and three dimensions. For the remaining number of errors (N_{R}), the mean of the horizontal absolute positions is calculated as $\bar{e}_{\text{excluding outliers}} = \frac{1}{N_{\mathrm{R}}} \sum_{i=1}^{N} e_i'$ A criterion for the establishing of correspondence should also be stated (e.g. allowing for correspondence to the closest position, correspondence on vertices or along lines). The criteria for finding the corresponding points shall be reported with the data quality evaluation result.
7	Parameter	Name: $e_{\max}$ Definition: is the threshold for accepted positional uncertainties Value type: Number
8	Value type	Measure
9	Value structure	–
10	Source reference	–
11	Example	–
12	Identifier	29

Enterprise Systems?



EOSDIS

Summary



“There are papers and web pages that describe the quality of my data.”

“Users increase our understanding of data quality. We need to keep them in the loop.”

“I use consistent types of Quality Measure across many products.”

“I use consistent Quality Measures across many products.”

“My metadata currently includes descriptions of the quality of my data.”

“My data quality information exists in databases or web services.”

“The quality of my data vary in time and space and different parameters have different quality measures and results.”

“I need to clearly and consistently explain how I measure quality.”

Acknowledgements

This work was partially supported by contract number NNG15HZ39C from NASA.

Raytheon

Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of NASA or The HDF Group.

The HDF Group